Received: 7 July 2018 | Revised: 7 May 2019 | Accepted: 8 May 2019

DOI: 10.1002/widm.1326

WIREs DATA MINING AND KNOWLEDGE DISCOVERY WILEY

ADVANCED REVIEW

Subgroup identification for precision medicine: A comparative review of 13 methods

Wei-Yin Loh | **Luxi Cao** | **Peigen Zhou**

Department of Statistics, University of Wisconsin, Madison, Wisconsin

Correspondence
Wei-Yin Loh, Department of Statistics, University of Wisconsin, Madison, WI.
Email: loh@stat.wisc.edu

Funding information
National Science Foundation, Grant/Award Number: DMS-1305725; University of Wisconsin Graduate School

Abstract

Natural heterogeneity in patient populations can make it very hard to develop treatments that benefit all patients. As a result, an important goal of precision medicine is identification of patient subgroups that respond to treatment at a much higher (or lower) rate than the population average. Despite there being many subgroup identification methods, there is no comprehensive comparative study of their statistical properties. We review 13 methods and use real-world and simulated data to compare the performance of their publicly available software using seven criteria: (a) bias in selection of subgroup variables, (b) probability of false discovery, (c) probability of identifying correct predictive variables, (d) bias in estimates of subgroup treatment effects, (e) expected subgroup size, (f) expected true treatment effect of subgroups, and (g) subgroup stability. The results show that many methods fare poorly on at least one criterion.

This article is categorized under:

Technologies > Machine Learning

Algorithmic Development > Hierarchies and Trees

Algorithmic Development > Statistics

Application Areas > Health Care

KEYWORDS

personalized medicine, prognostic variable, recursive partitioning, regression trees, tailored therapy

1 | INTRODUCTION

Because the effect of a treatment can vary substantially over a patient population, a central goal of precision medicine is identification of patient subgroups whose average response to a treatment is much higher or lower than the population average. To be useful, the subgroups should be defined in terms of biomarkers (such as laboratory test results, genetic profiles, and history and severity of illness) as well as demographic variables (such as age, gender, and race). A common approach in finding the subgroups is analysis of data from a randomized clinical trial. Following popular terminology, a variable is said to be "prognostic" if it conveys information on the likely outcome of a disease, independent of the treatment. Examples of such variables include patient age, family history of disease, disease stage, and prior therapy. A variable is said to be "predictive" if it identifies the likely benefit resulting from the treatment (Italiano, 2011). Predictive variables are also known as "treatment moderators" in some domains (Chen, Tian, Cai, & Yu, 2017). In statistical terms, a predictive variable has an interaction with the

treatment variable. A variable can be both prognostic and predictive. Methods for identifying subgroups often identify predictive variables as well.

There are few comparative studies of subgroup methods. Two studies compared some methods completed on one or two sets of data (Doove, Dusseldorp, Van Deun, & Van Mechelen, 2014; Lipkovich, Dmitrienko, & D'Agostino Sr., 2017). Another study used normally distributed simulated data (Alemayehu, Chen, & Markatou, 2017). The purpose of this article is to review 13 methods and compare their statistical properties and performance on seven criteria: (a) bias in selection of subgroup variables, (b) probability of false discovery, (c) probability of correctly identifying predictive variables, (d) bias in estimates of subgroup treatment effects, (e) expected true treatment effect of subgroups, (f) expected subgroup size, and (g) subgroup stability. The methods were selected because they have publicly available software that can be easily adapted for the simulation experiments. Because many of the methods are inapplicable to data with missing values in the predictor variables, the comparison is limited to completely observed data.

For the sake of brevity and simplicity, the methods are described for a binary response variable ($Y = 0, 1$) and a binary treatment variable ($Z = 0, 1$). Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ denote a vector of p predictor variables and let $(Y_i, Z_i, \mathbf{X}_i)$ denote the values taken by the ith observation. The methods here find subgroups with differential treatment effects, that is, the estimated treatment effects in the subgroup are larger (in absolute value) than those in its complement. Predictive variables are identified as those appearing in the definitions of the subgroups.

2 | SUBGROUP METHODS

2.1 | Tree methods

The model of Negassa, Ciampi, Abrahamowicz, Shapiro, and Boivin (2005) appears to be the earliest tree method for subgroup identification, but its software is not available.

1. **IT**: *Interaction trees* (Su, Tsai, Wang, Nickerson, & Bogong, 2009; Su, Zhou, Yan, Fan, & Yang, 2008). This algorithm quite faithfully follows the CART (classification and regression trees) approach (Breiman, Friedman, Olshen, & Stone, 1984). It recursively partitions the data with splits chosen to optimize an objective function and then prunes the resulting tree using the AIC (Akaike information criterion). Given a node t and a split s on variable X, let t_L and t_R denote the left and right subnodes of t. The split takes the form $s = \{X \le c\}$ for a constant c if X is ordinal, and $s = \{X \in A\}$ for a subset A of the levels of X if it is categorical. The value of c or A is that which maximizes the quantity

$$\frac{|(\bar{y}_{L1} - \bar{y}_{L0}) - (\bar{y}_{R1} - \bar{y}_{R0})|}{\hat{\sigma}\sqrt{n_{L0}^{-1} + n_{L1}^{-1} + n_{R0}^{-1} + n_{R1}^{-1}}} \tag{1}$$

   where $\bar{y}_{Lz}$, $\bar{y}_{Rz}$, n_{Lz} and n_{Rz} are the mean responses and sample sizes in t_L and t_R of the observations with treatment $Z = z$ ($z = 0, 1$), and $\hat{\sigma}$ is a pooled estimate of the error SD. This is equivalent to minimizing the p-value from testing the hypothesis that $\delta = 0$ in the model $Y = \eta + \beta Z + \gamma I(s) + \delta Z I(s) + \epsilon$. Although the variables appearing in the subgroups may be identified as predictive, the aggressive search for splits makes their identification unreliable. This is because variables that offer more ways to split a node have a higher chance to be chosen. Further, as shown later, maximizing quantity (1) produces biased estimates of treatment effects. The R functions in http://biopharmnet.com/subgroup-analysis-software/ were used to perform the computations here.
2. **SIDES**: *Subgroup identification based on differential effect search* (Lipkovich, Dmitrienko, Denne, & Enas, 2011). SIDES finds multiple alternative subgroups by identifying the best m splits of each node t that maximize a p-value criterion. In the examples and simulations below, we use $m = 2$ and the "differential effect splitting" p-value $2\{1 - \Phi(|T_L - T_R|/\sqrt{2})\}$, where T_L and T_R denote the test statistics for testing the one-sided hypothesis of treatment efficacy in t_L and t_R and Φ is the standard normal distribution function. For each split, the procedure is repeated on the subnode with the larger estimated treatment effect, up to a prespecified depth (we used `depth = 3` here). Heuristic adjustments are applied to the p-values to control for multiplicity of splits and correlations among the p-values. Once a variable is selected to split a node, it is not used to split subsequent nodes. As a result, SIDES cannot yield subgroups of the form $\{a < X \le b\}$ for finite values of a and b.
   SIDEScreen (Lipkovich & Dmitrienko, 2014) extends SIDES by adding a preliminary variable selection step. This is carried out by using SIDES to score the importance of the X variables first. Then those with high scores are applied to SIDES to find the subgroups. In the examples and simulations below, the "adaptive SIDEScreen" default is used, where a high score is

defined to be greater than 1 SD above the mean of the permutation null distribution of the maximum importance score. The software was obtained from http://biopharmnet.com/subgroup-analysis-software/. As with IT, the variables in the subgroups may be identified as predictive variables. But the multiplicity corrections for p-value adjustments do not completely remove selection bias.

3. *Virtual twins* (Foster, Taylor, & Ruberg, 2011) uses random forest (Breiman, 2001) to estimate the treatment effect $\tau(x) = P(Y = 1 \mid X = x, Z = 1) - P(Y = 1 \mid X = x, Z = 0)$ of each observation, with split variables $Z, X_1, \ldots, X_p$ and their interactions $ZX_1, \ldots, ZX_p$, and $(1 - Z)X_1, \ldots, (1 - Z)X_p$. Categorical variables are converted to dummy 0–1 variables. Then CART is applied to fit a classification or regression tree to the estimated $\tau(x)$ values to find the subgroups. If a classification tree is used, the two classes are defined by the estimated $\tau(x)$ being greater or less than a prespecified constant. If a regression tree is used, the subgroups are the terminal nodes with estimated treatment effect greater than $\varphi + 0.05$, where ϕ is the estimated marginal treatment effect of the whole training sample. The examples and simulations below used regression trees because classification trees often produced no subgroups. The trees were pruned with the default complexity parameter value of 0.02. (The alternative 0-SE pruning rule slightly reduced the probability of discovery under both null and non-null models.) Use of CART allows VT to be used for subgroup identification as well as identification of predictive variables, but the latter is unreliable due to the selection biases of CART (Loh, 2002; Loh & Shih, 1997). The R functions in http://biopharmnet.com/subgroup-analysis-software/ were used to perform the computations here.
4. **GUIDE**: *Generalized unbiased interaction detection and estimation* (Loh, 2002, 2009). GUIDE recursively partitions the data to form a binary tree whose terminal nodes define the subgroups. Here we consider only the *Gi option*, where at each node t, an interaction test is performed on each X variable to select one to split the data in the node into two subnodes (Loh, Fu, Man, Champion, & Yu, 2016; Loh, He, & Man, 2015; Loh, Man, & Wang, 2018). If X is a categorical variable with d levels $a_1, a_2, \ldots, a_d$ (missing categorical values are assigned their own level), the null hypothesis H_0: $Y = \tau Z + \sum_{k=1}^{d} \gamma_k I(X = a_k) + \epsilon$ is tested against the alternative hypothesis H_1: $Y = \sum_{k=1}^{d} \sum_{z=0}^{1} \delta_{kz} I(X = a_k, Z = z) + \epsilon$, where ϵ is assumed to be independent and normally distributed with mean 0 and constant variance. If X is an ordinal variable, it is first transformed into a categorical variable by bracketing its values at the node sample X-quartiles, with an additional level for missing values. The variable X with the smallest p-value from testing H_0 versus H_1 is selected to split the node. If X is categorical, the split takes the form $s = \{X \in A\}$, where A is a subset of the levels of X. If X is ordinal, the split takes the form $s_1 = \{X \le c \text{ or } X = \texttt{NA}\}$ or $s_2 = \{X \le c \text{ and } X \ne \texttt{NA}\}$, where `NA` denotes the missing value code. This approach to variable selection ensures that GUIDE does not have selection bias. Therefore it can be used for identification of subgroups and predictive variables.

   The selected values of A or c depend on the complexity of the linear models fitted in the subnodes. There are three choices: (a) *Gcon:* $EY = \beta_0 + \tau Z$, (b) *Glin:* $EY = \beta_0 + \beta X^* + \tau Z$, where X^* is the ordinal X variable yielding the smallest residual sum of squares, and (c) *Gstep:* $EY = \beta_0 + \sum_{X_j \in S} \beta_j X_j + \tau Z$, where S is the subset of variables yielding the smallest residual sum of square, obtained by stepwise linear regression. Categorical variables are included via their dummy 0–1 variables.

   Let S_L and S_R denote the residual sums of squares of the fitted models in the left and right subnodes t_L and t_R, respectively. The selected split (and the value of A or c) is the one that minimizes $S_L + S_R$. Partitioning continues recursively until the sample size in each node falls below a given threshold. Then the CART cross-validation (CV) pruning method is employed to reduce the size of the tree. Gcon and Glin employ the "0.5-SE rule," which gives the smallest subtree with CV estimate of mean squared error within 0.5 SE of the smallest CV estimate. Gstep uses the 0-SE rule. The software was obtained from http://pages.stat.wisc.edu/loh/guide.html.
5. **MOB**: *Model-based recursive partitioning* (Seibold, Zeileis, & Hothorn, 2016; Zeileis, Hothorn, & Hornik, 2008). MOB fits a parametric model (e.g., generalized linear model or Weibull accelerated failure time model) to the data in each node, with parameter values estimated as solutions to the score equations, the scores being partial derivatives of the log-likelihood. The variable selected to split a node is found by means of tests of independence between each X variable and the scores corresponding to the intercept and the treatment effect. Observations with missing values are excluded (Seibold, Zeileis, & Hothorn, 2017, Appendix 2). Given a prespecified level of statistical significance, Bonferroni adjustments are employed to determine whether any test is significant. If there is none, the node is not split. Otherwise, the variable with the smallest p-value is selected; the split point is chosen to minimize the sum of the negative log-likelihoods in the two subnodes.

   The examples and simulations below employ the `glmtree` function in the R package `partykit`. Because the objective here is to find subgroups defined by predictive variables (instead of prognostic variables), the `parm` option was used to restrict the independence tests to the scores for the treatment variable. There are two node models: (a) *MOBc* with $\log\{P$

$(Y=1)/P(Y=0)\} = \beta_0 + \tau Z$ and (b) *MOBm* with $\log\{P(Y=1)/P(Y=0)\} = \beta_0 + \sum_j \beta_j X_J + \tau Z$, where categorical variables are converted to dummy 0–1 variables. The trees are pruned with an AIC. MOBc, and MOBm can be used for identification of predictive variables but as shown below, MOBc has selection bias if there are prognostic variables.

2.2 | Nontree methods

1. **FindIt**: *Finding heterogeneous treatment effects* (Imai & Ratkovic, 2013). FindIt uses a penalized support vector machine to find predictive variables. Let $Y^* = (2Y-1)$ and let $\mathbf{U} = \left(X_1, X_2, \ldots, X_p, X_1^2, X_2^2, \ldots, X_p^2, X_1X_2, X_1X_3, \ldots, X_{p-1}X_p\right)$ be the vector consisting of all linear, quadratic, and two-factor interactions of the predictor variables. Let $\mathbf{V} = \left(ZX_1, \ldots, ZX_p, ZX_1^2, \ldots, ZX_p^2, ZX_1X_2, \ldots, ZX_{p-1}X_p\right)$ be the vector derived from $\boldsymbol{U}$ by multiplying its elements with Z. FindIt fits the support vector machine model $W(\mathbf{X}) = \mu + \beta\mathbf{U} + \gamma\mathbf{V}$, where W is a latent variable, with two LASSO penalties:

$$\left(\hat{\beta}, \hat{\gamma}\right) = \arg\min \sum_i \left|1 - Y_i^*(\mu + \beta\mathbf{U} + \gamma\mathbf{V})\right|_+^2 + \lambda_1 \sum_j |\beta_j| + \lambda_2 \sum_k |\gamma_k|.$$

The values of λ_1 and λ_2 are chosen by generalized CV. A variable X_i is considered to be predictive if at least one of ZX_i, ZX_i^2, $ZX_iX_1, ZX_iX_2, \ldots$ has a nonzero γ component. Let $\hat{W}$ be the fitted value and $\hat{W}^* = \min\left(\max\left(\hat{W}, -1\right), 1\right)$ be $\hat{W}$ truncated at ± 1. The estimated conditional average treatment effect at $\mathbf{X} = \mathbf{x}$ is $\hat{\tau}(\mathbf{x}) = \left\{\hat{W}^*(\mathbf{X}, Z=1) - \hat{W}^*(\mathbf{X}, Z=0)\right\}/2$ and the selected subgroup consists of the observations for which $\hat{\tau}(\mathbf{x}) > 0$. The R package *FindIt* (Egami, Ratkovic, & Imai, 2017) was used in the examples and simulations.
2. **ROWSi**: *Regularized outcome weighted subgroup identification* (Xu et al., 2015). Let $\pi = P(Z=1)$, $\xi(x) = \log(1+\exp(-x))$, and $\|\cdot\|$ denote the L_1 norm of a vector. Subgroups are defined by the sign of $\mathbf{X}'\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the minimizer of

$$n^{-1} \sum_{i=1}^{n} \frac{\xi\left(\{2Z_i - 1\}\mathbf{X}_i'\boldsymbol{\beta}Y_i\right)}{\pi(2Z_i - 1) + (1 - Z_i)} + \lambda_1 ||\boldsymbol{\beta}||_1 + \lambda_2 \eta(\boldsymbol{\beta})$$

and $\eta(\boldsymbol{\beta})$ is a penalty imposed on ordinal variables that take more than two values. The solution rests on many assumptions, including that $E(Y \mid \mathbf{X}, Z) = h(\mathbf{X}, (2Z-1)\mathbf{X}'\boldsymbol{\beta})$ for some unknown function h satisfying certain properties. The computations here used the R package *personalized* (Huling & Yu, 2018).
3. **PRIM**: *Patient rule induction method* (Chen, Zhong, Belousov, & Devanarayan, 2015). If Z is a binary treatment variable and Y is an uncensored continuous variable, the model fitted in each node is

$$\mathrm{EY} = \beta_0 + \beta_1 Z + \beta_2 \mathrm{ZI}(S) \tag{2}$$

where S denotes a subgroup. If Y is binary or right-censored, the left side of the model is replaced by the log odds ratio and log hazard ratio, respectively. Let $\hat{\beta}_i$ denote the estimated value of β_i $(i = 1, 2)$ and let $\bar{S}$ denote the complement of S. Assuming that treatment level $Z = 1$ has a negative effect on EY, permissible subgroups are required to satisfy some constraints, including: (a) the estimated treatment effect in S is less than that in $\bar{S}$ and (b) the statistical significance of the treatment effect in S is stronger than that in $\bar{S}$. Subgroups are found by splitting the training sample into two subsets and applying a bump-hunting procedure (Friedman & Fisher, 1999) to one subset with the p-value of the treatment effect as objective function. The other subset is used to pick the final subgroup from the pool of candidates. The computations here used the R package *SubgrpID* with the options `cv.iter` = 100 and `k.fold` = 5 (Huang et al., 2017).
4. **SeqBT**: *Sequential bootstrapping and aggregating of threshold from trees* (Huang et al., 2017). SeqBT uses the same model (2) as PRIM. The subgroup S consists of intersections of half-lines $\{X_j \le c_j\}$ or $\{X_j > c_j\}$ for some subset of predictor variables X_j, which is found iteratively. At each iteration, a search of the remaining X_j is carried out to find the value of c_j that optimizes the p-value for testing $\beta_2 = 0$ with the current S replaced with $S \cap \{X_j \le c_j\}$ and $S \cap \{X_j > c_j\}$; the X_j

with the smallest p-value is selected. A bootstrap step is included in the search for c_j. Iteration stops when the smallest p-value exceeds a prespecified threshold. The procedure is implemented in the R package *SubgrpID*.

5. **OWE**: *Outcome weighted estimation* (Chen et al., 2017). OWE is a general framework for subgroup identification using weighting or A-learning (Murphy, 2003). Assuming that the treatment variable $Z = \pm 1$ and given a loss function $M(y, v)$, which may be squared error or logistic loss, OWE employs the potential outcome approach of causal inference to find a score function $f(\mathbf{X})$ that minimizes the quantity

$$E\left\{\frac{M(Y, Zf(\mathbf{X}))}{Z\pi(\mathbf{X}) + (1-Z)/2} | \mathbf{X} = \mathbf{x}\right\}$$

where $\pi(\mathbf{X})$ is a propensity score which is known in randomized trials. The function f may be estimated with splines, additive models, or linear combinations of $\boldsymbol{X}$ (the last is used here). A lasso-type regularization penalty term may be added if the number of predictor variables is large. The subgroup with positive treatment effect ($T = 1$ vs. $T = -1$) consists of the observations with $f(\mathbf{x}) < 0$. The software is in the R package *personalized*.

3 | SIMULATION EXPERIMENTS

3.1 | Experimental design

Several simulation experiments were performed to evaluate the methods. Each experiment employed 2000 simulation trials with training samples of 400 observations per trial. Where permitted by the software, subgroups were required to have at least 30 observations with at least five for each treatment level. The response and treatment variables Y and Z were chosen to be binary so as to include as many methods as possible. Treatment assignment was independent of the covariates, mimicking randomized clinical trials. Because it is often known a priori in such trials that a nonzero treatment effect is either positive or negative, all the nonnull simulation models here had positive treatment effects.

While the identified subgroups in nontree methods are well-defined, being typically half-spaces, it is not clear in tree methods which terminal node (or union of terminal nodes) should be the identified subgroup. To reduce the number of potential subgroups, we defined a subgroup as *inadmissible* if it was the whole sample space (because it is not strictly a subgroup) or if its treatment effect estimate was not positive. An *admissible* subgroup is one that is not inadmissible. We chose the admissible subgroup with the largest positive estimated treatment effect as the identified subgroup in each trial. We did not consider the union of all subgroups with positive estimates of treatment effect because doing so reduces the average treatment effect of the union. Besides, the presumed use of the subgroup is to identify a target population for a future trial and a union of disjoint subgroups is harder to interpret than a single subgroup. True subgroup sizes and treatment effects were estimated with an independent test sample of 5,000 observations.

Ten predictor variables, $X_1, X_2, \ldots, X_{10}$, were employed. Their marginal distributions are given in Table 1, where $N(0, 1)$ denotes standard normal, $\texttt{Exp}(1)$ exponential with mean 1, $\texttt{Ber}(0.5)$ Bernoulli with success probability 0.5, and $M(10)$ multinomial with 10 equal-probability cells. All except the normally distributed X variables were mutually independent, and $\texttt{cor}(X_2, X_3) = \texttt{cor}(X_j, X_k) = 0.5$ for $j, k = 7, 8, 9, 10, \quad j \neq k$.

The Y variable was generated by the logit models shown in Table 2, which have the form

$$\texttt{logit} = \log\frac{P(Y=1)}{P(Y=0)} = f(x) + \theta I(Z=z)g(x).$$

Thus the *true* treatment effect of an observation with $X = x$ is

TABLE 1 Distributions of $X_1, X_2, \ldots, X_{10}$, and Z. All are mutually independent except $\texttt{cor}(X_2, X_3) = 0.5$ and $\texttt{cor}(X_j, X_k) = 0.5$ for $j, k = 7, 8, 9, 10, \quad j \neq k$

$X_1 \sim N(0, 1)$	$X_2 \sim N(0, 1)$	$X_3 \sim N(0, 1)$	$X_4 \sim \texttt{Exp}(1)$
$X_5 \sim \texttt{Ber}(0.5)$	$X_6 \sim M(10)$	$X_7 \sim N(0, 1)$	$X_8 \sim N(0, 1)$
$X_9 \sim N(0, 1)$	$X_{10} \sim N(0, 1)$	$Z \sim \texttt{Ber}(0.5)$	

TABLE 2 Three simulation models without treatment effect and eight with treatment effect

	Models without treatment effect	Prognostic	Predictive
B00	logit = 0	None	None
B01	logit = $0.5(X_1 + X_2)$	X_1, X_2	None
B02	logit = $0.5\left(X_1 + X_1^2 - 1\right)$	X_1	None
	Models with treatment effect		
B1	logit = $0.5(X_1 + X_2 - X_5) + 2ZI(X_6 = \text{odd})$	X_1, X_2, X_5	X_6
B2	logit = $0.5X_2 + 2ZI(X_1 > 0)$	X_2	X_1
B3	logit = $0.3(X_1 + X_2) + 2ZI(X_1 > 0)$	X_1, X_2	X_1
B4	logit = $0.3(X_2 + X_3 - 2) + 2ZX_4$	X_2, X_3	X_4
B5	logit = $0.2(X_1 + X_2 - 2) + 2ZI(X_1 < 1, X_6 = \text{odd})$	X_1, X_2	X_1, X_6
B6	logit = $0.5(X_2 - 1) + 2ZI(\mid X_1 \mid < 0.8)$	X_2	X_1
B7	logit = $0.2\left(X_2 + 2X_2^2 - 6\right) + 2ZI(X_1 > 0)$	X_2	X_1
B8	logit = $0.5X_2 + 2ZX_5$	X_2	X_5

$$\tau^*(x) = E(Y \mid X = x, Z = 1) - E(Y \mid X = x, Z = 0)$$

$$= P(Y = 1 \mid X = x, Z = 1) - P(Y = 1 \mid X = x, Z = 0)$$

$$= \frac{\exp\{f(x) + \theta g(x)\}}{1 + \exp\{f(x) + \theta g(x)\}} - \frac{\exp\{f(x)\}}{1 + \exp\{f(x)\}}.$$

Given a subgroup G, the *true* treatment effect τ_G^* is estimated by the mean of $\tau^*(x)$ among the test observations in G. The *estimated* treatment effect $\hat{\tau}_G$ is obtained from the training observations in G as follows.

Gcon, IT and SIDES. These three methods fit the linear model $E(Y) = \beta_0 + \beta_1 I(Z = z)$ in each node and $\hat{\tau}_G$ is the least-squares estimate of β_1 for the training data in G.

Glin. This fits a simple linear model $E(Y) = \beta_0 + \beta^* X^* + \theta I(Z = z)$ to each node where X^* is the best linear predictor in the node. $\hat{\tau}_G$ is the least-squares estimate of θ for the training data in G.

Gstep. This fits a stepwise linear model $E(Y) = \beta_0 + \sum_j \beta_j X_j + \theta I(Z = z)$ to each node and $\hat{\tau}_G$ is the least-squares estimate of θ for the training data in G.

MOBc, PRIM, and SeqBT. These fit the logistic model $\log\{P(Y = 1)/(Y = 0)\} = \beta_0 + \theta I(Z = z)$ in each node and

$$\hat{\tau}_G = \frac{\exp\left(\hat{\beta}_0 + \hat{\theta}\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\theta}\right)} - \frac{\exp\left(\hat{\beta}_0\right)}{1 + \exp\left(\hat{\beta}_0\right)},$$

where $\hat{\beta}_0$ and $\hat{\theta}$ are the MLEs of β_0 and θ, respectively, for the training data in G.

MOBm. This fits the multiple linear logistic model $\log\{P(Y = 1)/(Y = 0)\} = \beta_0 + \sum_j \beta_j X_j + \theta I(Z = z)$ in each node and

$$\hat{\tau}_G = \frac{\exp\left(\hat{\beta}_0 + \sum_j \hat{\beta}_j X_j + \hat{\theta}\right)}{1 + \exp\left(\hat{\beta}_0 + \sum_j \hat{\beta}_j X_j + \hat{\theta}\right)} - \frac{\exp\left(\hat{\beta}_0 + \sum_j \hat{\beta}_j X_j\right)}{1 + \exp\left(\hat{\beta}_0 + \sum_j \hat{\beta}_j X_j\right)},$$

where $\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\theta}\right)$ are the MLEs of $(\beta_0, \beta_1, \ldots, \theta)$ for the training data in G.

ROWSI, OWE, and VT. $\hat{\tau}_G$ is the difference between the training sample means of Y for the two treatment groups in G.

FindIt. $\hat{\tau}_G$ is the mean of the values of $\hat{\tau}(\mathbf{x})$ (defined in earlier description of FindIt) among the training observations in G.

3.2 | Results for models B00–B02

3.2.1 | Variable selection bias

Table 3 gives estimates of the probabilities that $X_1, \ldots, X_{10}$ are selected by each method in models B00, B01 and B02. Figure 1 shows plots of the values. For tree methods, the values are the probabilities that each variable is selected to split the root node of a tree, before it is pruned (for SIDES, they are the probabilities that each variable is selected to split the root node using parameter values `width = depth = 1`). For nontree methods, they are the probabilities that each variable has a nonzero interaction with Z; frequencies of multiple variables with nonzero coefficients are divided equally, except for dummy variables from the same categorical variable, which are counted only once. A method has unbiased variable selection if its probabilities are all equal to 0.10. The results show that only Gcon, Glin and Gstep are unbiased—their selection probabilities are all within two simulation standard errors of 0.10. The other methods have varying degrees of bias. IT and VT are the worst. They are heavily biased against the binary variable X_5 and in favor of the 10-level categorical variable X_6; these are properties inherited from CART (see Loh, 2002; Loh & Shih, 1997). OWE and ROWSi are also biased toward X_6, although not for the same reason. MOBc is unbiased under B00, but it is not under B01 and B02 where it is biased toward the prognostic variables (X_1 and X_2 in B01 and X_1 in B02). MOBm is biased under B02, because it tends to pick up the quadratic prognostic effect of X_1. SIDES is biased against binary (X_5) and categorical (X_6) variables. PRIM and SeqBT are biased against the binary variable X_5. FindIt is biased toward the exponential variable X_4 and the categorical variable X_6.

3.2.2 | Probability of false discovery

Table 4 gives the probabilities of false subgroup discovery (Type I error) of the methods under B00, B01, and B02. They are estimated by the proportions of simulation trials yielding admissible subgroups. The results, presented graphically in Figure 2, show that PRIM, ROWSi, SeqBT, and OWE have the largest probabilities of error (from 0.27 to 0.62). VT, SIDES, and FindIt form the middle group, with probabilities of error ranging from 0.13 to 0.17. The methods with best control of probability of Type I error are, in order, IT, Gstep, MOBm, Glin, MOBc, and Gcon.

3.3 | Results for models B1–B8

3.3.1 | Probability of selecting a predictive variable

Figure 3 plots the probability that each method correctly selects the predictive variable in models B1–B8. For tree methods, this is the probability that the predictive variable (or variables in the case of B5) is selected to split the root node of the tree, before it is pruned (for SIDES, it is the probability that a variable is selected to split the root node, using parameter values `width = depth = 1`). For nontree methods, it is the frequency that the estimated regression coefficient of the predictive variable is nonzero. The results show that IT, VT, SeqBT, MOBc, MOBm, Gcon, Glin, and Gstep are most likely to select the right predictive variables; ROWSi, OWE, and FindIt are the least. SIDES and PRIM are in the middle, with probabilities between 0.50 and 0.80.

3.3.2 | Mean subgroup size

Figure 4 plots the mean size of the subgroups for each method, conditional on a subgroup being found, where size is measured by the proportion of test observations in the subgroup. The results show that FindIt, OWE and ROWSi tend to have the largest subgroup size (at least 80%), followed by PRIM. VT tends to yield the smallest subgroups.

3.3.3 | True subgroup treatment effect

The large mean subgroup sizes of FindIt, OWE, and ROWSi are offset by their relatively small treatment effect sizes, as shown in Figure 5 which plots the median true effects (estimated from the test samples) of their subgroups. VT, MOBc, Gcon, Glin, Gstep, and IT have consistently the largest true treatment effects.

TABLE 3 Variable selection probabilities for models without treatment effect; simulation SEs approximately 0.0067

Method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Model B00: logit = 0										
IT	0.106	0.081	0.088	0.090	0.003	0.285	0.079	0.091	0.093	0.082
Gcon	0.103	0.093	0.087	0.110	0.104	0.112	0.102	0.095	0.093	0.101
Glin	0.101	0.098	0.089	0.113	0.099	0.106	0.107	0.098	0.091	0.098
Gstep	0.097	0.098	0.087	0.111	0.104	0.110	0.102	0.096	0.094	0.101
MOBc	0.114	0.107	0.102	0.115	0.107	0.095	0.102	0.085	0.088	0.086
MOBm	0.108	0.106	0.107	0.113	0.113	0.081	0.102	0.083	0.091	0.096
SeqBT	0.118	0.114	0.121	0.117	0.026	0.105	0.104	0.095	0.099	0.100
PRIM	0.128	0.117	0.130	0.126	0.006	0.088	0.102	0.097	0.102	0.105
FindIt	0.095	0.089	0.092	0.138	0.091	0.176	0.068	0.087	0.079	0.086
ROWSi	0.098	0.097	0.094	0.090	0.096	0.149	0.092	0.094	0.096	0.094
OWE	0.097	0.095	0.099	0.086	0.087	0.172	0.089	0.092	0.094	0.089
SIDES	0.111	0.118	0.102	0.108	0.045	0.054	0.108	0.143	0.102	0.108
VT	0.048	0.046	0.040	0.028	0.001	0.642	0.050	0.050	0.048	0.048
Model B01: logit = $0.5(X_1 + X_2)$										
IT	0.078	0.068	0.084	0.099	0.001	0.300	0.084	0.094	0.103	0.088
Gcon	0.105	0.091	0.084	0.115	0.111	0.096	0.106	0.111	0.091	0.090
Glin	0.101	0.092	0.082	0.122	0.113	0.102	0.107	0.103	0.086	0.092
Gstep	0.099	0.091	0.090	0.123	0.105	0.092	0.115	0.099	0.09	0.096
MOBc	0.267	0.235	0.088	0.068	0.066	0.051	0.057	0.064	0.053	0.051
MOBm	0.104	0.093	0.108	0.110	0.112	0.096	0.100	0.093	0.084	0.100
SeqBT	0.101	0.109	0.118	0.107	0.030	0.130	0.098	0.101	0.101	0.106
PRIM	0.118	0.111	0.116	0.111	0.004	0.105	0.110	0.102	0.116	0.107
FindIt	0.090	0.090	0.087	0.163	0.077	0.169	0.082	0.079	0.082	0.081
ROWSi	0.097	0.096	0.094	0.091	0.097	0.150	0.098	0.092	0.092	0.092
OWE	0.091	0.082	0.088	0.072	0.079	0.247	0.087	0.087	0.085	0.084
SIDES	0.151	0.100	0.100	0.100	0.061	0.022	0.133	0.115	0.111	0.108
VT	0.082	0.076	0.046	0.022	0	0.610	0.046	0.036	0.042	0.038
Model B02: logit = $0.5\left(X_1 + X_1^2 - 1\right)$										
IT	0.063	0.103	0.087	0.105	0.005	0.280	0.090	0.087	0.078	0.103
Gcon	0.113	0.094	0.088	0.093	0.114	0.095	0.097	0.101	0.109	0.096
Glin	0.111	0.097	0.091	0.093	0.117	0.091	0.094	0.101!	0.108	0.097
Gstep	0.109	0.099	0.094	0.093	0.115	0.095	0.095	0.100	0.105	0.095
MOBc	0.378	0.074	0.073	0.078	0.073	0.071	0.064	0.063	0.064	0.061
MOBm	0.157	0.092	0.102	0.118	0.096	0.088	0.080	0.092	0.089	0.086
SeqBT	0.102	0.117	0.107	0.115	0.025	0.126	0.101	0.108	0.103	0.097
PRIM	0.115	0.108	0.115	0.129	0.006	0.090	0.119	0.102	0.105	0.110
FindIt	0.098	0.086	0.082	0.137	0.084	0.181	0.082	0.079	0.087	0.083
ROWSi	0.095	0.097	0.094	0.092	0.095	0.152	0.092	0.097	0.094	0.093
OWE	0.089	0.087	0.085	0.078	0.079	0.223	0.087	0.091	0.088	0.093
SIDES	0.118	0.131	0.093	0.137	0.053	0.069	0.084	0.087	0.097	0.131
VT	0.093	0.042	0.03	0.029	0	0.621	0.05	0.045	0.042	0.046

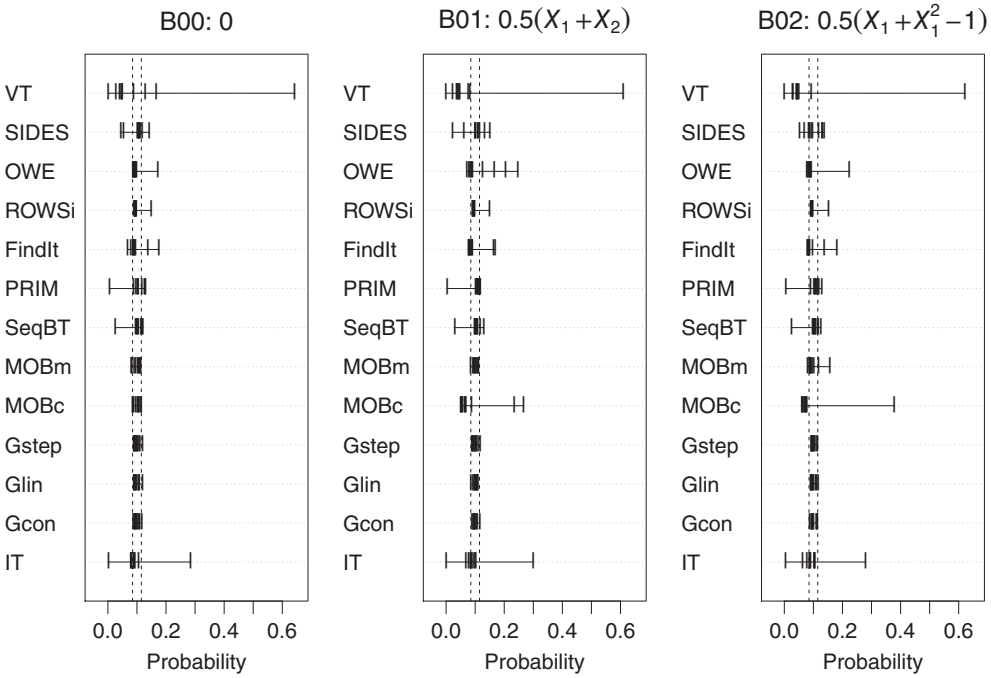

FIGURE 1 Plots of variable selection frequencies in Table 3. Each frequency value is marked with a short vertical bar; horizontal lines connect the smallest and largest selection frequencies for each method; dashed vertical lines mark two simulation standard errors around unbiasedness level of 0.10

TABLE 4 Estimated probabilities of false discovery (Type I error)

Method	B00	B01	B02	Method	B00	B01	B02
Gcon	0.016	0.110	0.106	FindIt	0.126	0.141	0.172
Glin	0.012	0.079	0.061	ROWSi	0.481	0.497	0.483
Gstep	0.015	0.009	0.022	OWE	0.427	0.274	0.292
MOBc	0.041	0.129	0.099	SIDES	0.140	0.136	0.148
MOBm	0.022	0.026	0.028	VT	0.150	0.145	0.152
SeqBT	0.342	0.430	0.374	IT	0	0.001	0
PRIM	0.623	0.602	0.619				

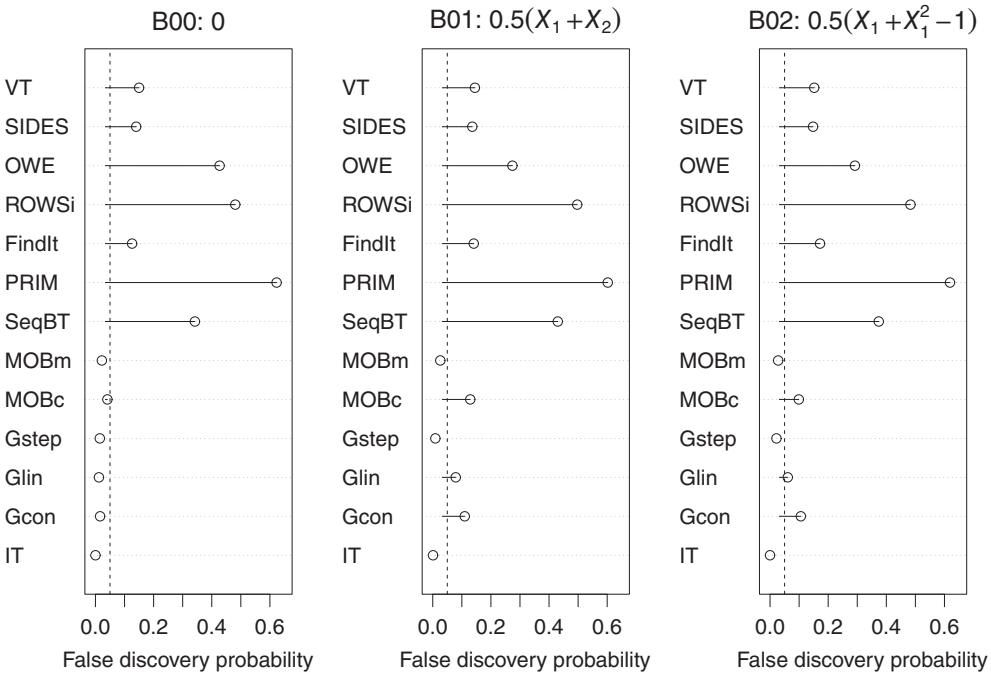

FIGURE 2 Plots of probability of false discovery (Type I error). For Gcon, MOBc and VT, the probabilities are upper bounds. Vertical dotted lines mark the 0.05 level

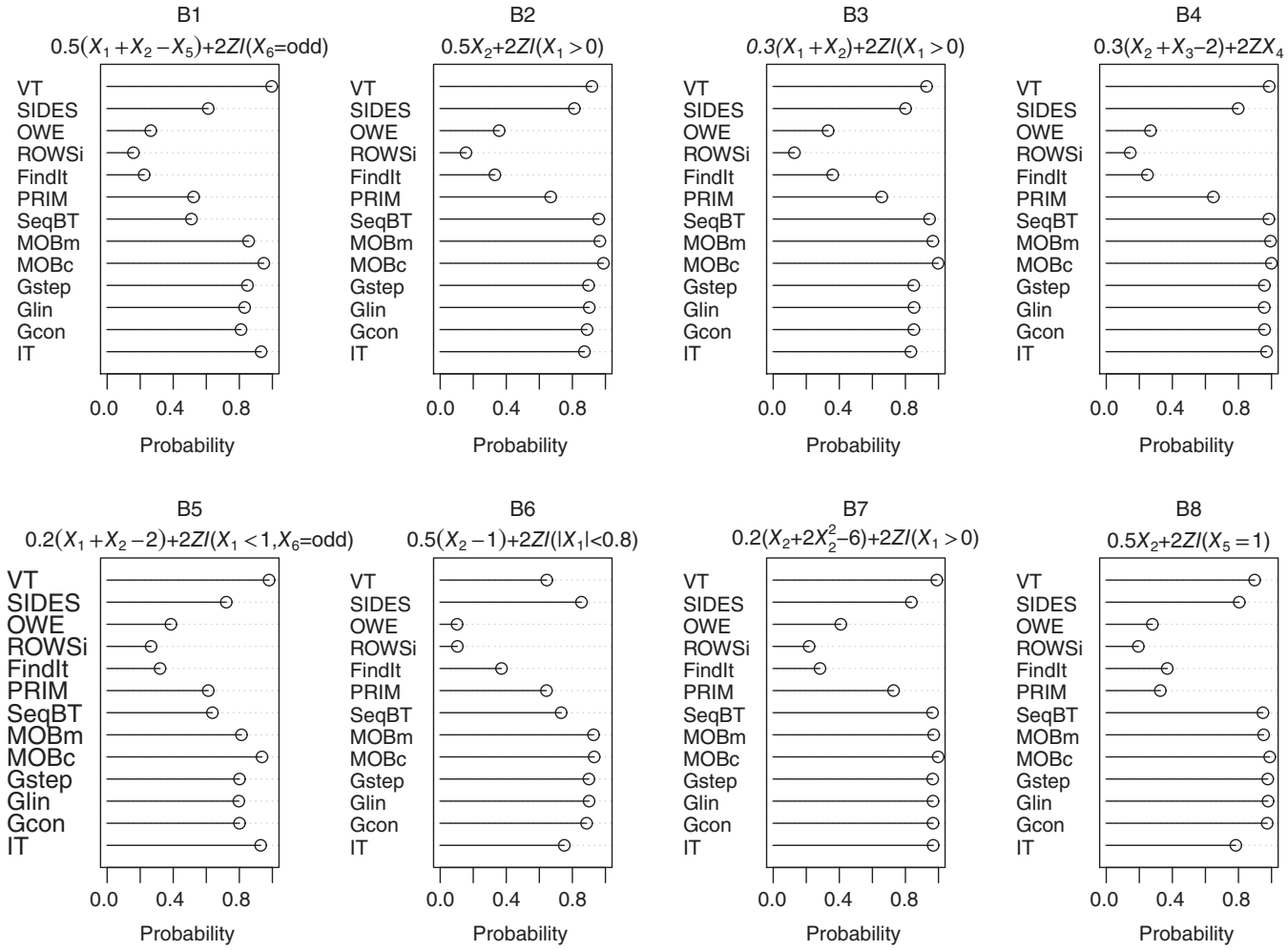

FIGURE 3 Probability of selecting a predictive variable at the first split for tree methods. For nontree methods, it is the probability that a predictive variable is among the selected variables

3.3.4 | Treatment effect bias

Some methods over-estimate the treatment effect in the selected subgroups. For each simulation trial i yielding a selected subgroup G_i, let S_i and S_i^* denote the sets of training and test observations belonging to G_i. Let $\hat{\tau}_i$ and $\hat{\tau}_i^*$ denote the estimated average treatment effects in G_i computed from S_i and S_i^*, respectively. The relative bias is estimated by the median of $\left(\hat{\tau}_i - \hat{\tau}_i^*\right)/\hat{\tau}_i^*$ over the simulation trials that yield subgroups (the median is used instead of the mean because $\hat{\tau}_i^*$ may be very small). The results, shown in Figure 6, reveal that SIDES, IT, PRIM, and SeqBT tend to have the largest relative bias—their subgroup treatment effect estimates are 20–50% larger than the true treatment effects. OWE and ROWSi have essentially no bias and FindIt is almost the same, except in model B6 where it has a large negative bias. MOBc, Gcon, and Glin have the next smallest relative bias. (The estimated biases are inevitably slightly overstated because the selected subgroup is required by design to have positive treatment effect.)

4 | REAL DATA

The simulations only show aggregate properties of the methods. To reveal features of individual subgroups, the methods were applied to three real data sets, which were originally collected to estimate overall treatment effects.

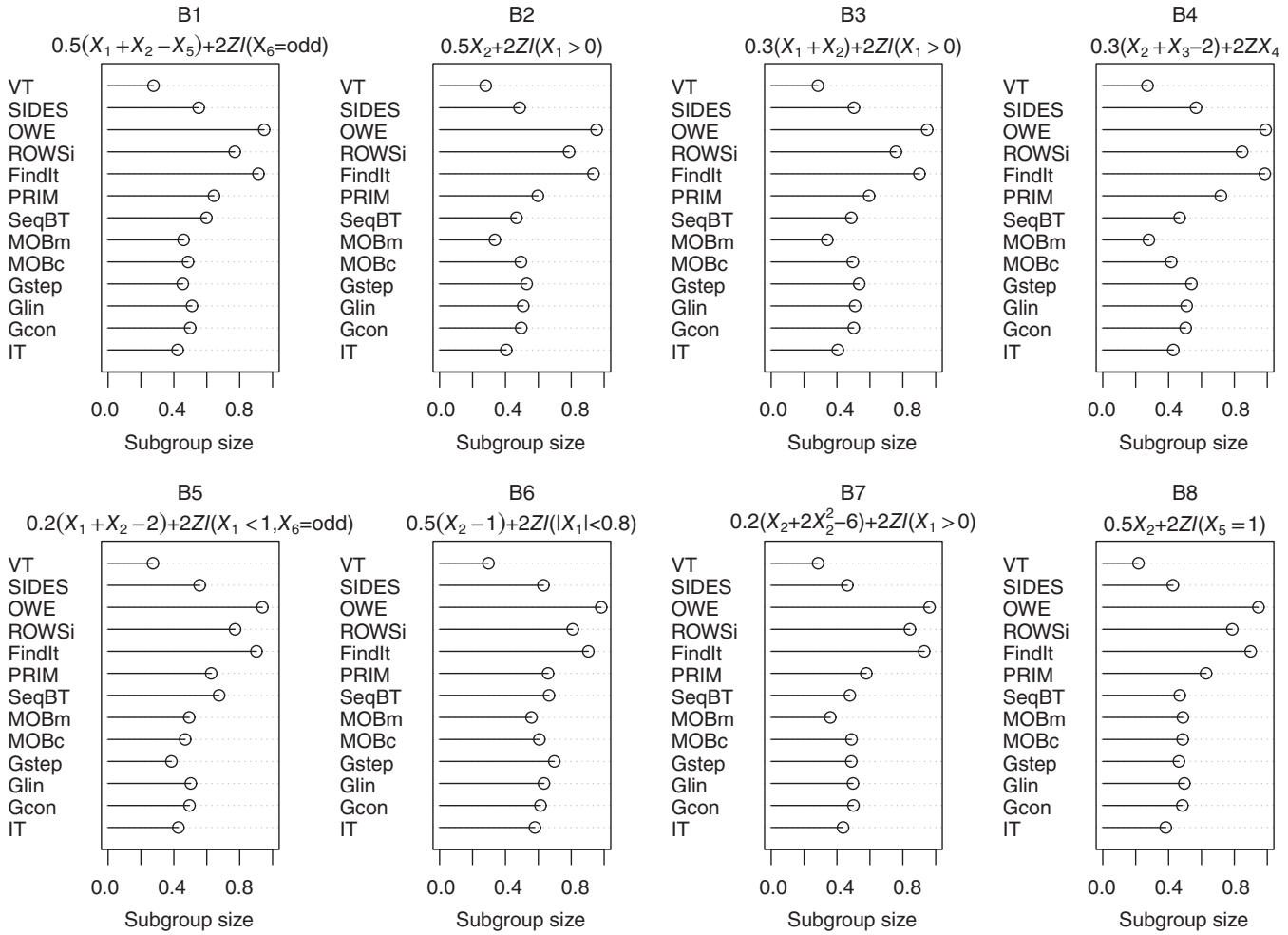

FIGURE 4 Conditional mean subgroup size as proportion of test observations

4.1 | Work training

The data consist of observations from 722 workers in a national supported work program (Dehejia & Wahba, 1999; LaLonde, 1986). A random sample of 297 disadvantaged workers (such as welfare recipients, ex-addicts, ex-offenders, and young school dropouts) were assigned to a training program (treatment) while the others served as controls. The response variable was binary (1 if 1978 earnings were greater than that in 1975; 0 otherwise). Predictor variables were `u75` (1 if unemployed in 1975; 0 otherwise), `age` (17–55), `educ` (years of education), `race` (white, black, Hispanic), 1975 earnings, college degree (yes or no), and `marr` (1 for married, 0 for unmarried). Imai and Ratkovic (2013) and Egami et al. (2017) used the data to identify subgroups of workers for whom the training program was beneficial.

Table 5 gives the results. Gcon, Gstep, IT, and MOBm detected no subgroups. Glin found a subgroup consisting of married workers, MOBc a subgroup defined by the unemployment variable `u75`, and SeqBT and SIDES a subgroup defined by `race`. PRIM found a larger subgroup defined by `race`, `educ`, and `age`. FindIt, ROWSi, and OWE found subgroups defined by linear combinations of all the predictor variables. VT produced random subgroups, due to the inherent randomness of random forest. Estimates of the treatment effects in the subgroups that were found ranged from 0.08 to 0.23, with subgroups sizes from 117 to 646. Overall, the results are rather inconclusive, because there is little consistency among methods. The subgroups defined by linear combinations of variables are hard to interpret.

4.2 | Breast cancer

The data are from a randomized trial of 686 subjects with primary node positive breast cancer (Schumacher et al., 1994). Treatment was hormone therapy (`horTh`) versus no therapy and the response was recurrence-free survival time in days, with

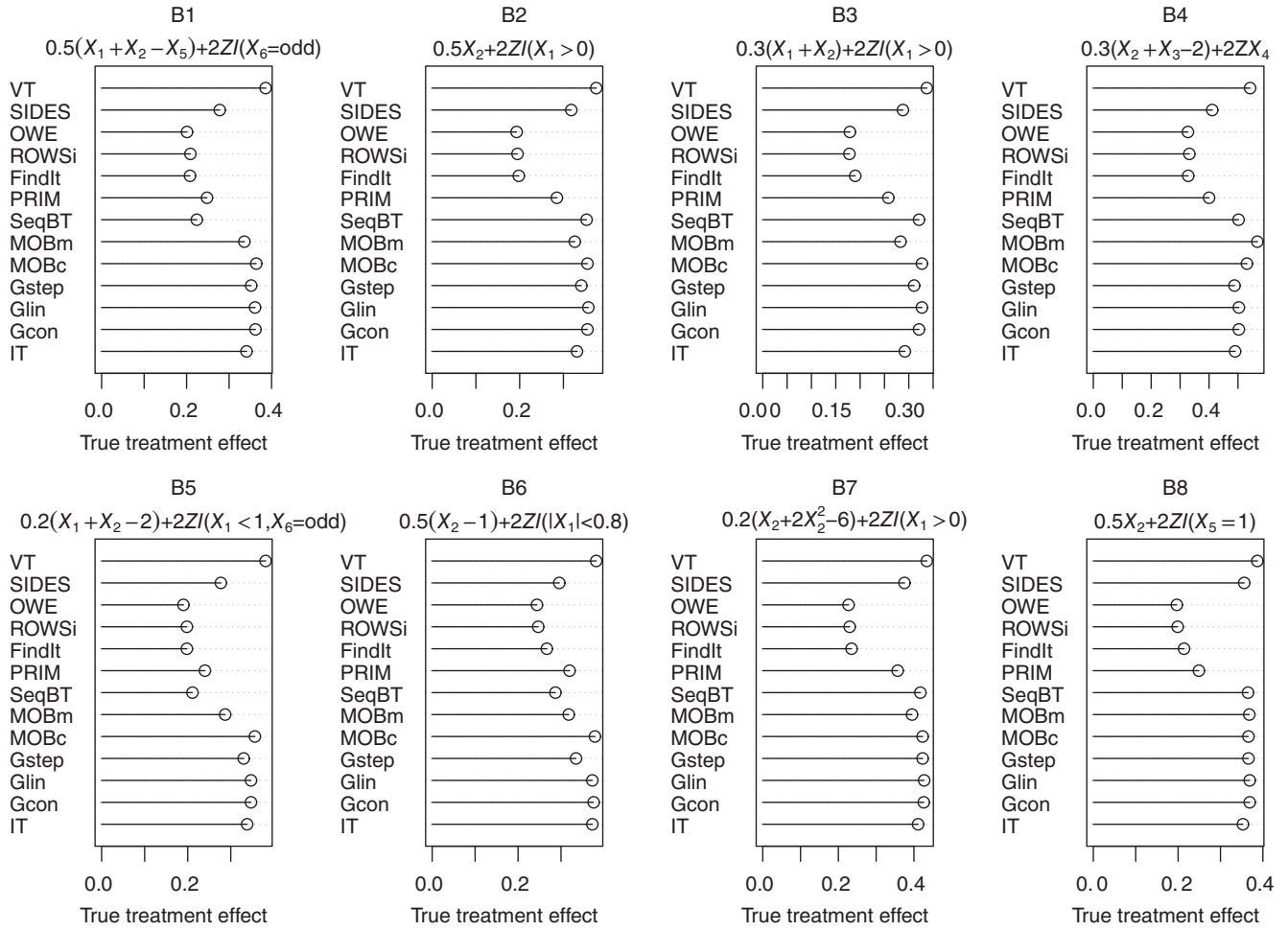

FIGURE 5 Conditional median true treatment effect of subgroups

56% censoring. Predictor variables were `age` (21–80), `tsize` (tumor size, 3–120 mm), `pnodes` (number of positive lymph nodes, 1–51), `progrec` (progesterone receptor status, 0–2,380 fmol), `estrec` (estrogen receptor status, 0–1,144 fmol), `menostat` (menopausal status, pre vs. post), and `tgrade` (tumor grade, 1–3). The data were previously used by Loh et al. (2015, 2018) to find subgroups with differential treatment effects. They found that `progrec` and `estrec` were predictive variables and `pnodes` was a prognostic variable.

FindIt, Gstep, ROWSi, and VT were excluded here because they are inapplicable to censored response data. Table 6 gives the results for the other methods. Gcon, MOBc, and SeqBT identified `progrec`, SIDES found `estrec`, and PRIM found `pnodes`. Glin, IT, MOBm, and OWE did not find any subgroups. There were large variations in the subgroup sizes and their estimated treatment effects. Not surprisingly, large treatment effects were associated with small subgroups. Figures 7–10 show the Kaplan–Meier survival curves in the subgroups and their complements. The plots for SIDES and PRIM show, somewhat counterintuitively, that there were subgroups where hormone therapy was worse than no therapy. This is likely due to overfitting of differential treatment effects between subgroups.

4.3 | Heart failure

The data are from two randomized studies of left ventricular dysfunction on the efficacy of enalapril, an angiotensin-converting enzyme inhibitor, on mortality and hospitalization for heart failure (SOLVD Investigators, 1991). The SOLVD-T trial enrolled 2,569 subjects with history of overt congestive heart failure and the SOLVD-P trial enrolled 4,228 subjects without history of overt congestive heart failure. The response variable was survival time from enrollment to death or hospitalization. Table 7 lists the predictor variables.

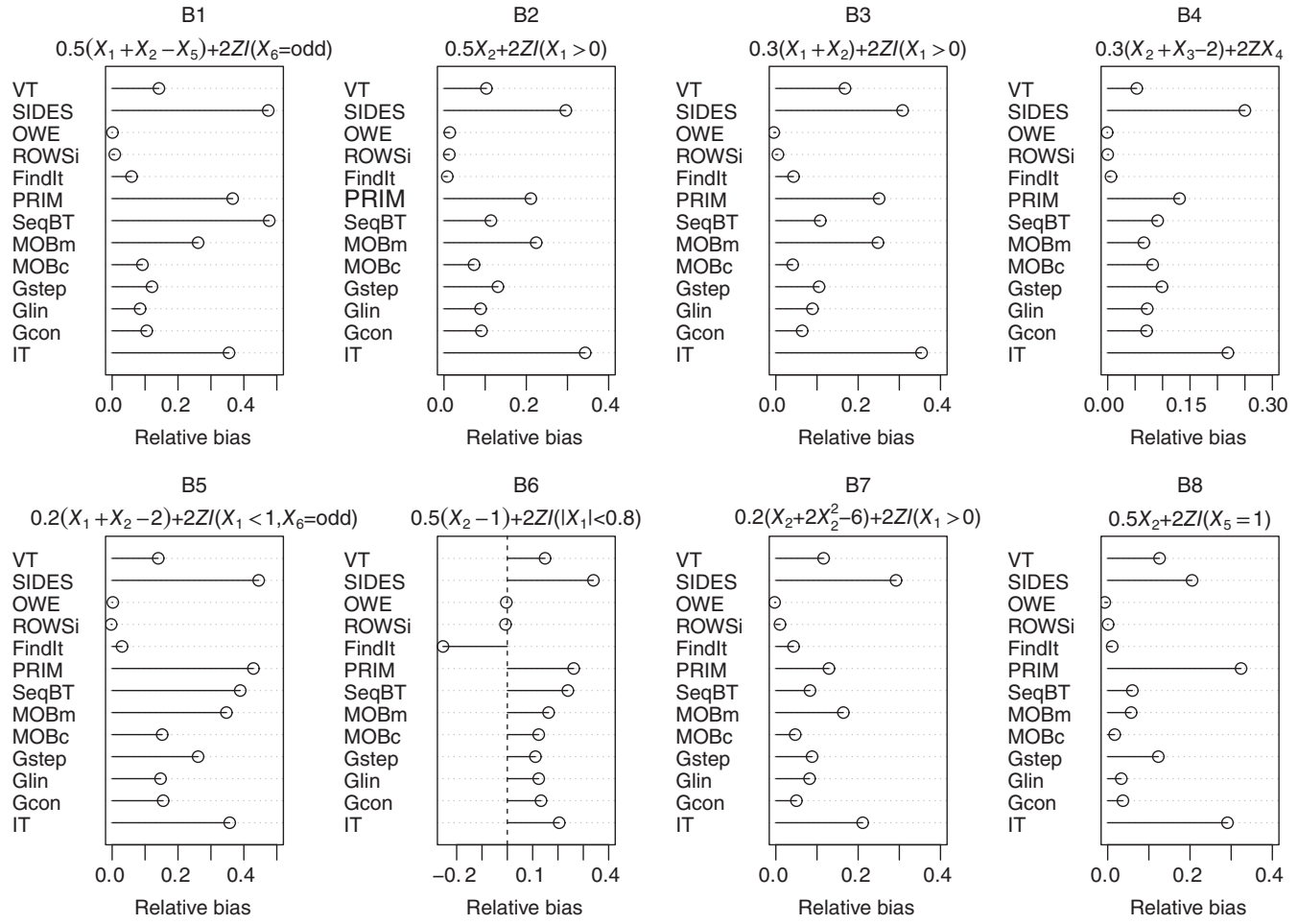

FIGURE 6 Conditional median relative bias of estimated treatment effects

TABLE 5 Subgroups, their number of observations, and estimated treatment effects for work training data

Method	Subgroup	Number of observations	Effect
Gcon, Gstep, IT, MOBm	None	0	0
Glin	`marr` = 1	117	0.23
MOBc	`u75` = 1	289	0.08
PRIM	`educ` ≥9 & `age` > 18 & `race` = `black` or `white`	479	0.11
SeqBT, SIDES	`race` = `black` or `white`	646	0.11
FindIt	Linear combination of all variables	558	0.11
OWE	Linear combination of all variables	497	0.14
ROWSi	Linear combination of all variables	409	0.18
VT	Random	Random	Random

The original purpose of the studies was to see if the treatment had an overall beneficial effect on survival. We used the data here to look for subgroups with differential treatment effects. FindIt, Gstep, ROWSi, and VT were again excluded because they are inapplicable to censored response data. The Gcon tree in Figure 11 shows that the subgroup, `lvef` ≤26 and `crackles` =1, has the largest estimated treatment effect. Glin, MOBm, PRIM, and SeqBT also found `lvef` to be predictive but MOBc and SIDES identified `copd` and other variables. Glin additionally found `nyha` to be the best linear prognostic predictor. IT and OWE found no subgroups. Table 8 gives the results for all the methods. Because the treatment is expected to have a positive effect on survival (negative effect on hazard risk), the table lists only subgroups with large negative treatment

Method	Subgroup	Number of observations	Effect
Gcon, MOBc	progrec >21	405	−0.66
PRIM	pnodes ≤16	657	−0.42
SeqBT	progrec >65 & pnodes <9	238	−1.20
SIDES	estrec >0	604	−0.49
Glin, IT, MOBm, OWE	None	0	0
FindIt, Gstep, ROWSi, VT	Inapplicable to censored data		

TABLE 6 Subgroups, their number of observations, and estimated treatment effects (in terms of log hazard) for breast cancer data

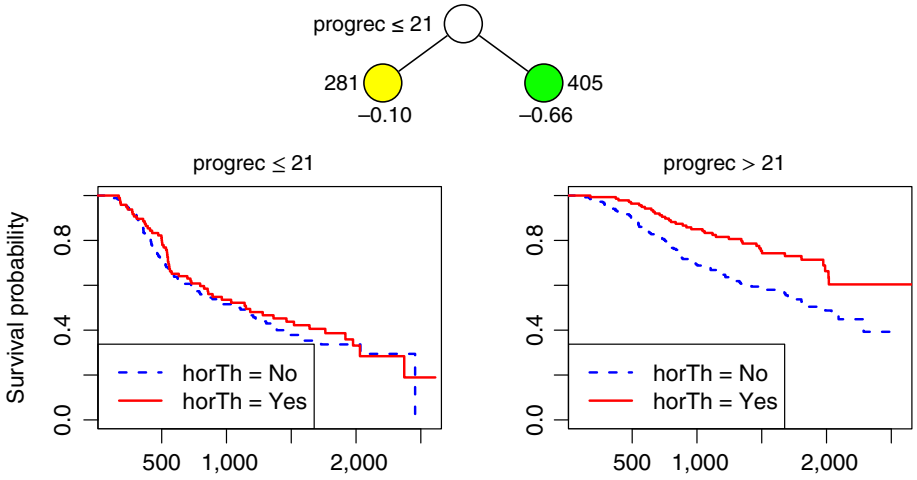

FIGURE 7 Gcon subgroup (in green) for breast cancer data; sample sizes and estimated treatment effects (log relative risks) beside and below nodes

effects. Figures 12–17 show the survival curves in the subgroups of the other methods. Again, as in the previous example, Figure 16 shows that treatment is worse than no treatment in the complementary subgroup for SIDES.

5 | CONCLUSIONS

Unlike other machine learning applications where the sole goal is accurate prediction of future observations, a subgroup identification method needs to satisfy multiple criteria in order to be useful. This paper employed publicly available and simulated data to compare 13 methods with regard to their biases in variable selection and treatment effect estimation, probability of

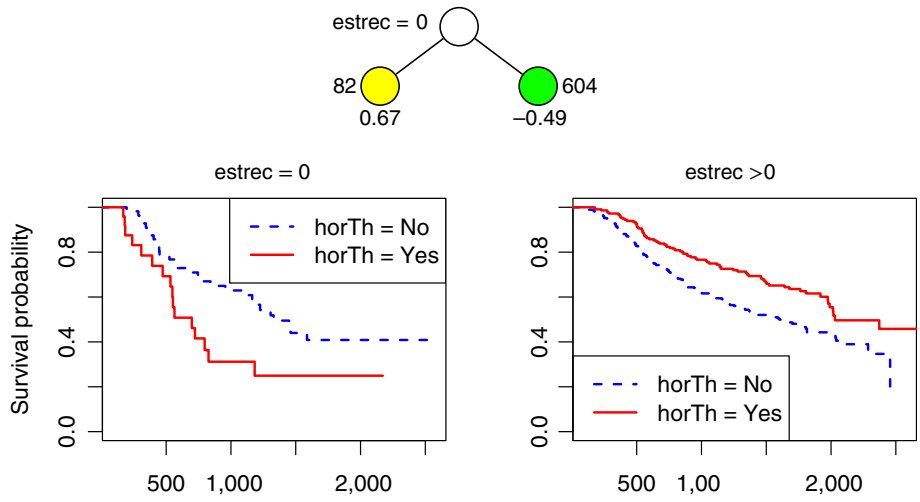

FIGURE 8 SIDES subgroup (in green) for breast cancer data; sample sizes and estimated treatment effects (log relative risks) beside and below nodes

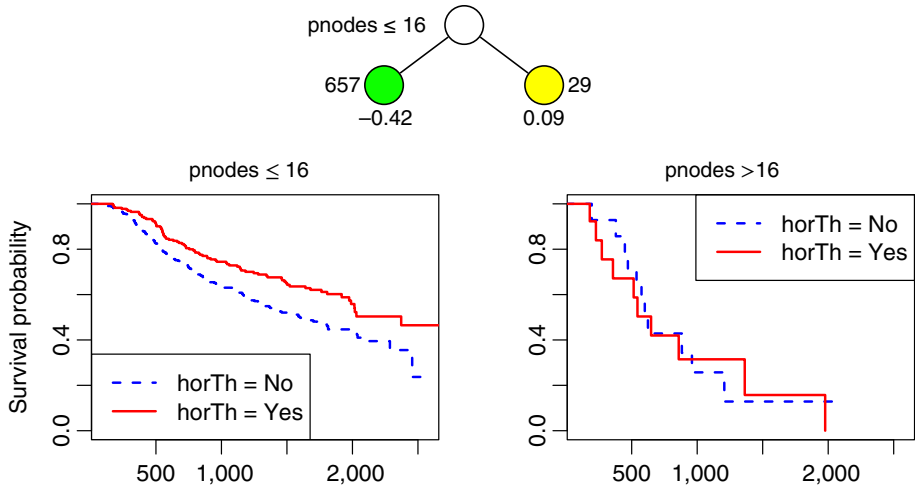

FIGURE 9 PRIM subgroup (in green) for breast cancer data; sample sizes and estimated treatment effects (log relative risks) beside and below nodes

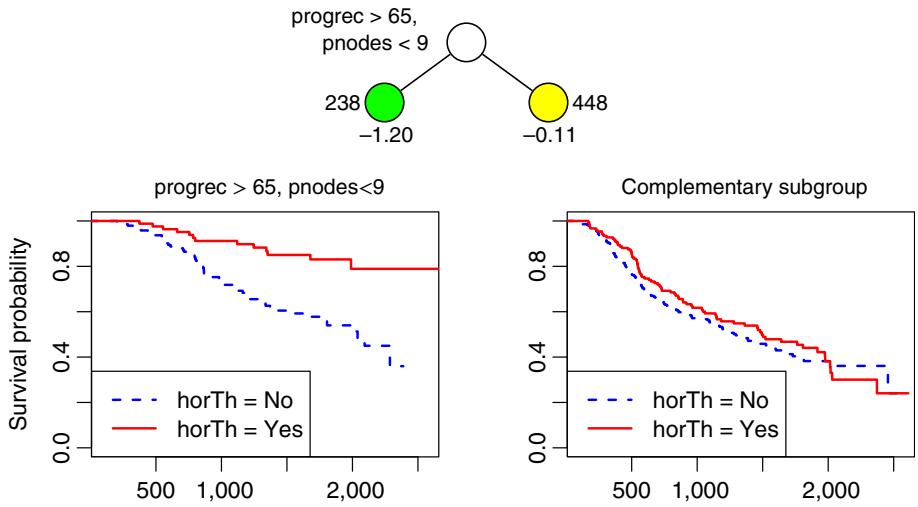

FIGURE 10 SeqBT subgroup (in green) for breast cancer data; sample sizes and estimated treatment effects (log relative risks) beside and below nodes

false discovery, probability of selecting the correct predictive variable, mean subgroup size, true mean treatment effect in the subgroup, and bias in the treatment effect estimates.

In terms of selection bias, VT, IT, and MOBc perform most poorly, the first two due to their adoption of the CART exhaustive search paradigm and the latter due to its inability to separate the effects of prognostic variables. The methods with

TABLE 7 Predictor variables in heart data

Name	Description	Name	Description
`trt`	Treatment vs. placebo	`weightkg`	Weight in kg (45–136)
`study`	SOLVD-P vs SOLVD-T	`anydiurbl`	Binary (0, 1)
`age`	Age (27–80)	`avsys`	Continuous (85–180)
`avdia`	Continuous (50–110)	`sodium`	Serum sodium (129–149)
`creatinine`	Serum creatinine (0.4–3.5)	`copd`	Presence of COPD (0, 1)
`depedema`	Binary (0, 1)	`histk`	Binary (0, 1)
`diabet`	Diabetic status (0, 1)	`beat`	Heart rate (45–120)
`crackles`	Binary (0, 1)	`gender`	Gender
`smoke`	Smoking status (current, former or never)	`himi`	History of myocardial infarction (0, 1)
`lvef`	Left ventricular ejection fraction (10–35)	`nyha`	New York Heart Association functional class (1–4)

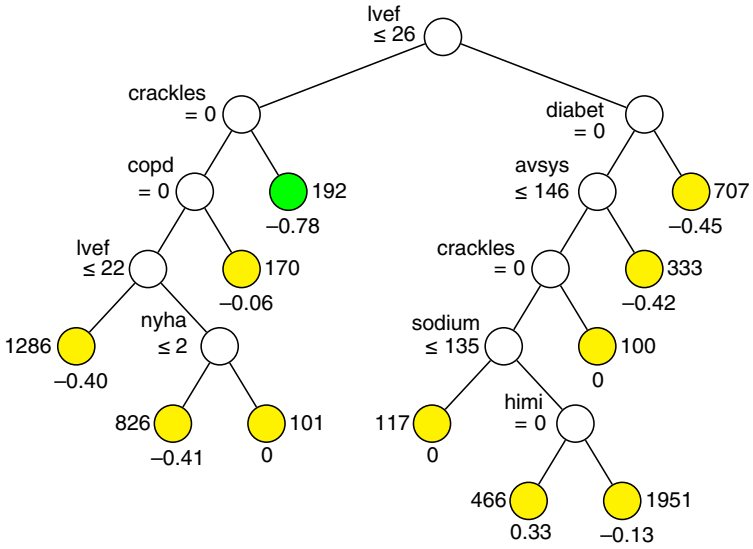

FIGURE 11 Gcon tree for heart data. Sample size and treatment effect (log relative risk of treated vs. untreated) printed beside and below each node. Node with selected subgroup is in green color

Method	Subgroup	Number of observations	Effect
Gcon	`lvef` ≤26 & `crackles` = 1	192	−0.78
Glin	`lvef` ≤26 with `nyha` linear prognostic predictor	2,575	−0.41
MOBc	`anydiurbl` = 1 & `copd` = 0	2,387	−0.47
MOBm	`lvef` >28 & `diabet` = 1	571	−0.83
PRIM	`lvef` ≤28	3,204	−0.38
SIDES	`beat` ≤110 & `avsys` ≤176 & `copd` = 0	5,736	−0.31
SeqBT	`lvef` ≤29 & `sodium` >140 & `age` < 72	1,271	−0.55
IT, OWE	None	0	0
Others	Inapplicable to censored data		

TABLE 8 Subgroups and their number of observations, and estimated treatment effect (in terms of log hazard) for heart data

best control of variable selection bias are Gcon, Glin, and Gstep (Figure 1). In terms of probability of false discovery, the worst method is PRIM, with a probability consistently above 0.50. It is followed by ROWSi, SeqBT, OWE, SIDES, and FindIt, roughly in that order. The methods with best control of the probability are, in order, IT, Gstep, MOBm, Glin, MOBc, and Gcon (see Figure 2), although this seems to come at a price for IT, which found no subgroups in all three data sets. In terms of probability of selecting the correct predictive variable, the poorest methods are ROWSi, FindIt, OWE, PRIM, and SIDES, in that order. The other methods have fairly high probabilities (see Figure 3). In terms of subgroup size, OWE, ROWSi, and FindIt tend to produce the largest subgroups, although large subgroups typically are associated with small true

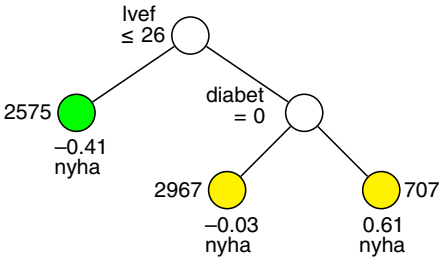

FIGURE 12 Glin tree for heart data. Sample size printed beside node and treatment effect (log relative risk of treated vs. untreated) and name of linear prognostic variable printed below node. Node with selected subgroup is in green color

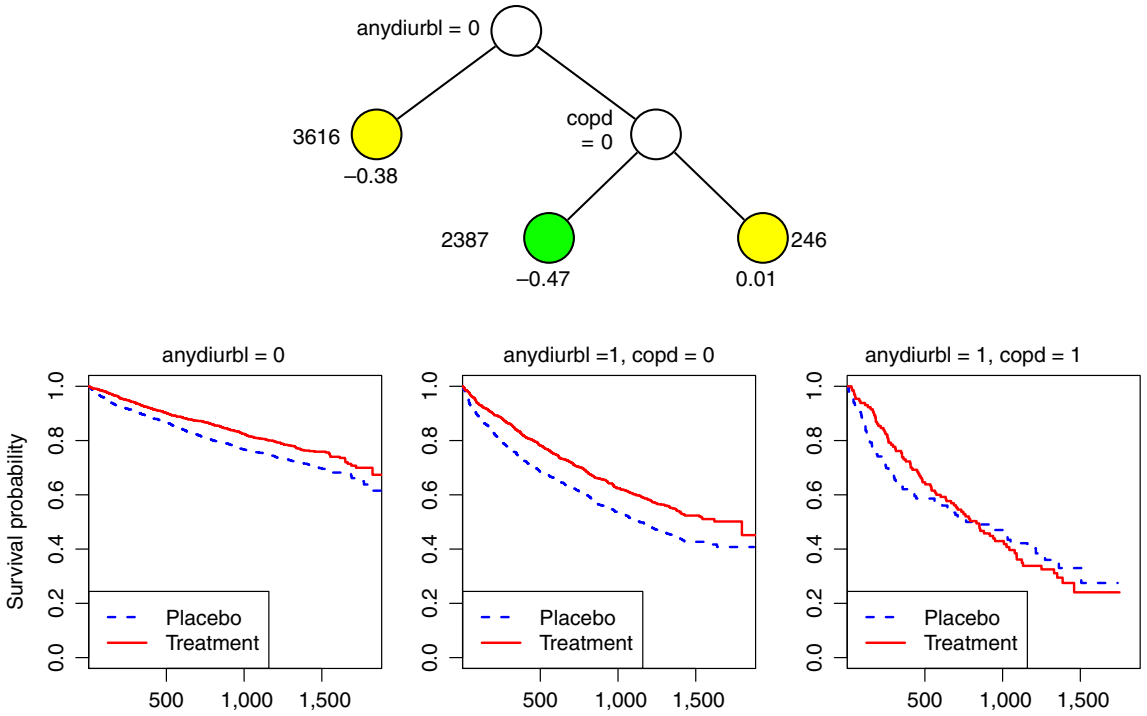

FIGURE 13 MOBc tree for heart data. Sample size and treatment effect (log relative risk of treated vs. untreated) printed beside and below each node. Node with selected subgroup is in green color

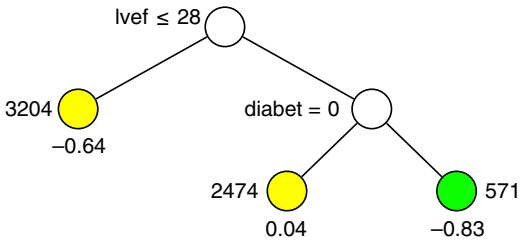

FIGURE 14 MOBm tree for heart data. Sample size and treatment effect (log relative risk of treated vs. untreated) printed beside and below each node. Node with selected subgroup is in green color

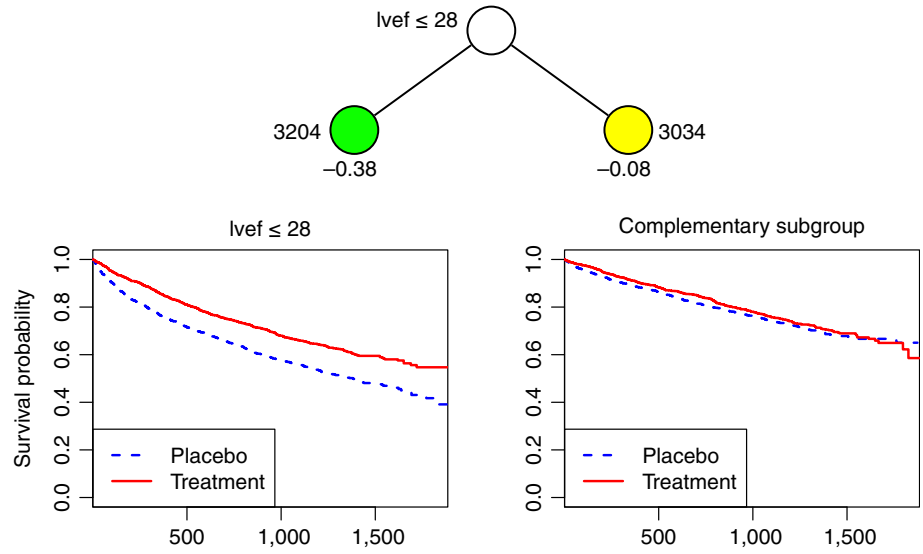

FIGURE 15 Survival curves of PRIM subgroup (in green) and its complement for heart data. Sample size and treatment effect (log relative risk of treated vs. untreated) printed beside and below each node

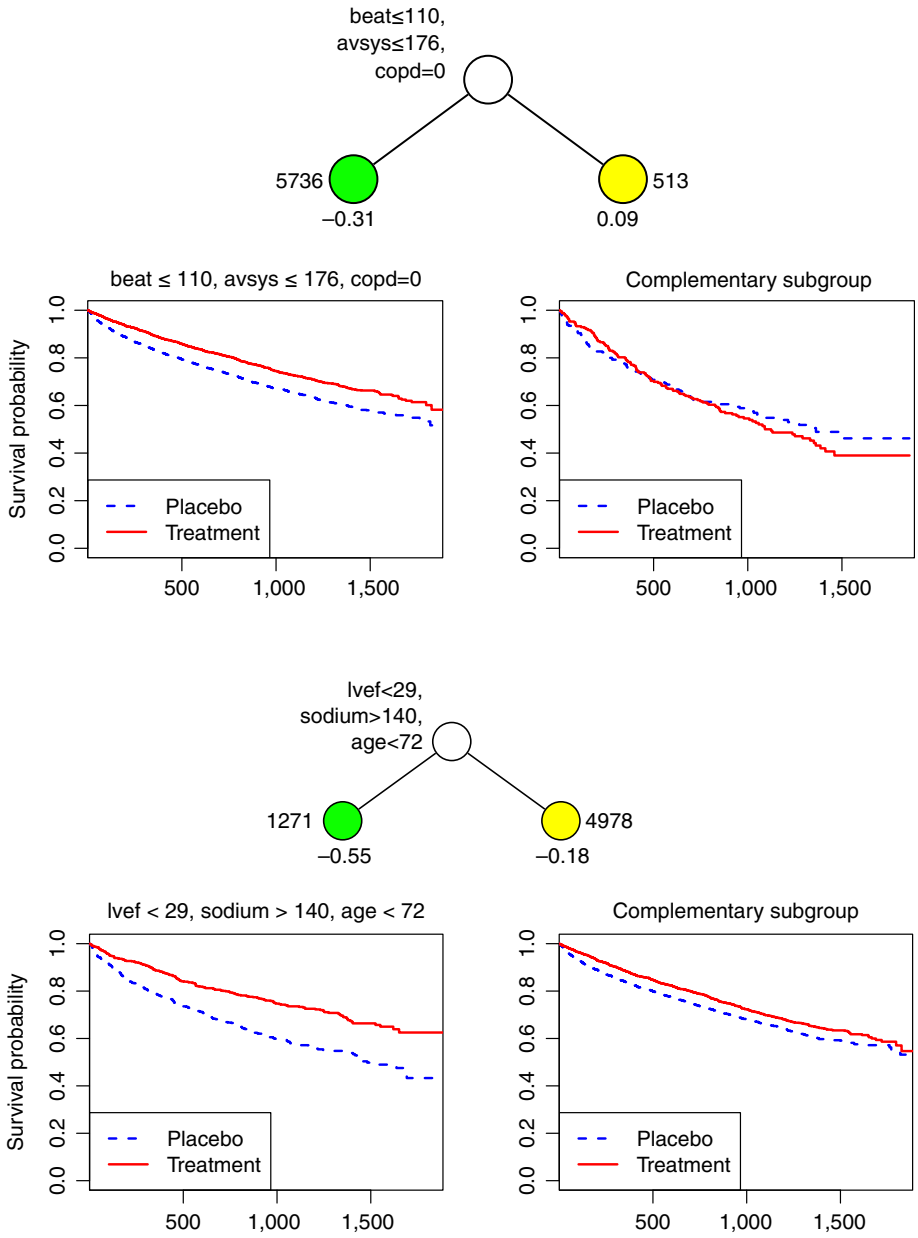

FIGURE 16 Survival curves of SIDES subgroup (in green) and its complement for heart data. Sample size and treatment effect (log relative risk of treated vs. untreated) printed beside and below each node

FIGURE 17 Survival curves of SeqBT subgroup (in green) and its complement for heart data. Sample size and treatment effect (log relative risk of treated vs. untreated) printed beside and below each node

TABLE 9 Types of response variables and ability to accept missing data values

	Response variable type			Missing
Method	**Binary**	**Continuous**	**Censored**	**values**
FindIt	Yes	Yes	No	No
GUIDE	Yes	Yes	Yes	Yes
IT	Yes	Yes	Yes	Yes
MOB	Yes	Yes	Yes	No
OWE	Yes	Yes	Yes	No
PRIM	Yes	Yes	Yes	No
ROWSi	Yes	No	No	No
SeqBT	Yes	Yes	Yes	No
SIDES	Yes	Yes	Yes	Yes
VT	Yes	Yes	No	No

TABLE 10 Summary of properties

	FI	GC	GL	GS	IT	MC	MM	OW	PR	RO	SB	SI	VT
Variable bias	□	✓	✓	✓	×	×	□	□	□	□	□	□	×
P(false discovery)	□	✓	✓	✓	✓	✓	✓	×	×	×	×	□	□
Predictive var. ID	×	✓	✓	✓	✓	✓	✓	×	□	×	□	□	✓
Effect bias	□	□	□	□	×	□	×	✓	×	✓	×	×	□
Interpretability	×	✓	✓	✓	✓	✓	✓	×	✓	×	✓	✓	✓
Stability	✓	✓	✓	✓	✓	✓	✓	×	×	×	×	□	×
Missing values	×	✓	✓	✓	×	×	×	×	×	×	×	□	×

FI, FindIt; GC, Gcon; GL, Glin; Gs, Gstep; MC, MOBc; MM, MOBm; OW, OWE; PR, PRIM; RO, ROWSi; SB, SeqBT; SI, SIDES. "Variable bias" refers to bias in variable selection when there is no treatment effect (Figure 1). "P(false discovery)" refers to probability of Type I error (Figure 2). "Predictive var. ID" refers to probability of identifying predictive variables (Figure 3). "Effect bias" refers to bias of subgroup treatment effect estimates (Figure 6). "Interpretability" refers to ease of interpretation of the subgroups; subgroups defined by linear combinations of variables are difficult to interpret. "Stability" refers to randomness of subgroups. "Missing values" refers to the ability of the method to accept missing values. For each criterion, the methods are divided into three groups. A checkmark (✓) is given to those in the top group, a cross (×) to those in the bottom group, and a square (□) to those in the middle group that are satisfactory but not the best.

subgroup treatment effects (see Figures 4 and 5). Some methods yield overly optimistic estimates of treatment effects. Chief among them are IT, SIDES, PRIM, and SeqBT. The methods with the least-biased treatment effect estimates are OWE, ROWSi, MOBc, Glin, and Gcon (see Figure 6). Awareness of these properties is helpful for choosing among different methods, as the three real examples demonstrate.

In real applications, methods that use CV for parameter tuning or tree pruning produce random subgroups unless the random seed is locked. Gcon, Glin, Gstep, and FindIt lock the seed in the software. OWE, PRIM, ROWSi, SeqBT, and VT let the user change the seed or base the seed on the computer clock. As a result, their subgroups are random and hence unstable. (The seed can be fixed by the user of these algorithms, but this opens the door to "cheating," where a user tries different seeds until he obtains a satisfactory result.) IT, MOBc, and MOBm use AIC or BIC (Bayesian information criterion) for pruning and hence are stable. Because SIDES uses resampling-based Bonferroni corrections, the results are theoretically random; but the effect is not as apparent as in CV.

Only completely observed data were used because 8 of the 13 methods (FindIt, MOBc, MOBm, OWE, PRIM, ROWSi, SeqBT, and VT) do not accept missing values. In evaluating a predictor variable X for split selection, IT and SIDES exclude observations with missing values in X (Lipkovich et al., 2017, Sec. 10); this approach is known to induce selection bias in CART (Kim & Loh, 2001). Gcon, Glin, and Gstep use all observations.

Tables 9 and 10 summarize the properties of the methods. In our opinion, the most important for practical applications are unbiased variable selection (to reduce the chance of mis-identifying subgroups and predictive variables), unbiased estimates of subgroup treatment effects (to avoid over optimism), and probability of false discovery. The ability to accept data with missing values is a plus but also often a necessity. Based on the simulation and publicly available data results here, the GUIDE methods are among the best, if not the best.

ACKNOWLEDGMENTS

The authors thank two referees for their helpful comments. They also thank Xiaogang Su, Ilya Lipkovich, Achim Zeileis, and Menggang Yu for assistance with the IT, SIDES, MOB, ROWSi, and OWE software. The first author is grateful to Hock Peng Chan for arranging his visit to the National University of Singapore where the manuscript was completed. W.-Y.L.'s research was supported in part by NSF grant DMS-1305725 and a grant from the University of Wisconsin Graduate School.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

RELATED WIREs ARTICLES

Classification and regression trees

ORCID


Wei-Yin Loh https://orcid.org/0000-0001-6983-2495

REFERENCES

Alemayehu, D., Chen, Y., & Markatou, M. (2017). A comparative study of subgroup identification methods for differential treatment effect: Performance metrics and recommendations. *Statistical Methods in Medical Research*, *27*, 3658–3678. https://doi.org/10.1177/0962280217710570

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Chen, G., Zhong, H., Belousov, A., & Devanarayan, V. (2015). A PRIM approach to predictive-signature development for patient stratification. *Statistics in Medicine*, *34*, 317–342.

Chen, S., Tian, L., Cai, T., & Yu, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*, *73*, 199–1209.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*, 1053–1062.

Doove, L. L., Dusseldorp, E., Van Deun, K., & Van Mechelen, I. (2014). A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment-subgroup interactions. *Advances in Data Analysis and Classification*, *8*, 403–425.

Egami, N., Ratkovic, M., & Imai, K. (2017). *Findit: Finding heterogeneous treatment effects [Computer software manual].* (R package version 1.1.2). Retrieved from https://CRAN.R-project.org/package=FindIt

Foster, J. C., Taylor, J. M. G., & Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, *30*, 2867–2880.

Friedman, J. H., & Fisher, N. I. (1999). Bump hunting in high-dimensional data. *Statistics and Computing*, *9*, 123–143.

Huang, X., Sun, Y., Trow, P., Chatterjee, S., Chakravartty, A., Tian, L., & Devanarayan, V. (2017). Patient subgroup identification for clinical drug development. *Statistics in Medicine*, *36*, 1414–1428.

Huling, J. D., & Yu, M. (2018). *Subgroup identification using the personalized package.* Retrieved from https://arxiv.org/abs/1809.07905

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *Annals of Applied Statistics*, *7*, 443–470.

Italiano, A. (2011). Prognostic or predictive? It's time to get back to definitions! *Journal of Clinical Oncology*, *29*, 4718–4719.

Kim, H., & Loh, W.-Y. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, *96*, 589–604.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, *76*, 604–620.

Lipkovich, I., & Dmitrienko, A. (2014). Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES. *Journal of Biopharmaceutical Statistics*, *24*, 130–153.

Lipkovich, I., Dmitrienko, A., & D'Agostino, R. B., Sr. (2017). Tutorial in biostatistics: datadriven subgroup identification and analysis in clinical trials. *Statistics in Medicine*, *36*(1), 136–196.

Lipkovich, I., Dmitrienko, A., Denne, J., & Enas, G. (2011). Subgroup identification based on differential effect search—A recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, *30*, 2601–2621.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, *12*, 361–386.

Loh, W.-Y. (2009). Improving the precision of classification trees. *Annals of Applied Statistics*, *3*, 1710–1737.

Loh, W.-Y., Fu, H., Man, M., Champion, V., & Yu, M. (2016). Identification of subgroups with differential treatment effects for longitudinal and multiresponse variables. *Statistics in Medicine*, *35*, 4837–4855.

Loh, W.-Y., He, X., & Man, M. (2015). A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine*, *34*, 1818–1833.

Loh, W.-Y., Man, M., & Wang, S. (2018). Subgroups from regression trees with adjustment for prognostic effects and post-selection inference. *Statistics in Medicine*, *38*, 545–557. https://doi.org/10.1002/sim.7677

Loh, W.-Y., & Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica*, *7*, 815–840.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, Series B*, *65*, 331–355.

Negassa, A., Ciampi, A., Abrahamowicz, M., Shapiro, S., & Boivin, J. R. (2005). Tree-structured subgroup analysis for censored survival data: validation of computationally inexpensive model selection criteria. *Statistics and Computing*, *15*, 231–239.

Schumacher, M., Baster, G., Bojar, H., Hübner, K., Olschewski, M., Sauerbrei, W., … Rauschecker, H. F. (1994). Randomized 2 × 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology*, *12*, 2086–2093.

Seibold, H., Zeileis, A., & Hothorn, T. (2016). Model-based recursive partitioning for subgroup analyses. *International Journal of Biostatistics*, *12*, 45–63.

Seibold, H., Zeileis, A., & Hothorn, T. (2017). Individual treatment effect prediction for amyotrophic lateral sclerosis patients. *Statistical Methods in Medical Research*, *27*, 3104–3125. https://doi.org/10.1177/0962280217693034

SOLVD Investigators. (1991). Effect of Enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. *New England Journal of Medicine*, *325*(5), 293–302.

Su, X., Tsai, C. L., Wang, H., Nickerson, D. M., & Bogong, L. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, *10*, 141–158.

Su, X., Zhou, T., Yan, X., Fan, J., & Yang, S. (2008). Interaction trees with censored survival data. *International Journal of Biostatistics*, *4*, Article 2.

Xu, Y., Yu, M., Zhao, Y.-Q., Li, Q., Wang, S., & Shao, J. (2015). Regularized outcome weighted subgroup identification for differential treatment effects. *Biometrics*, *71*, 645–653.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514.

How to cite this article: Loh W-Y, Cao L, Zhou P. Subgroup identification for precision medicine: A comparative review of 13 methods. *WIREs Data Mining Knowl Discov*. 2019;9:e1326. https://doi.org/10.1002/widm.1326